DOCUMENT RESUME

ED 161 929 TM 007 956

AUTHOR Downing, Steven M.; Mehrens, William A.
TITLE Six Single-Administration Reliability Coefficients for Criterion-Referenced Tests: A Comparative Study.
PUB DATE [Mar 78]
NOTE 12p.; Paper presented at the Annuel Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Comparative Analysis; *Comparative Statistics; *Criterion Referenced Tests; Factor Analysis; Statistical Analysis; *Test Reliability
IDENTIFIERS Test Theory

ABSTRACT

Four criterion-referenced reliability coefficicents were compared to the Kuder-Richardson estimates and to each other. The Kuder-Richardson formulas 20 and 21, the Livingston, the Subkoviak and two Huynh coefficients were computed for a random sample of 33 criterion-referenced tests. The Subkoviak coefficient yielded the highest mean value; Huynh's Kappa yielded the lowest. The two Huynh coefficients were highly positively correlated with the Kuder-Richardson 20 and 21 coefficients, and with each other; the Livingston and the Subkoviak indexes were highly correlated with each other. A two-factor principle components analysis suggested that the Subkoviak coefficient measured a test characteristic that differed from the classical internal-consistency coefficients. (Author/CTM)

Six Single-Administration Reliability Coefficients for Criterion-Referenced Tests: A Comparative Study

Steven M. Downing

William A. Mehrens

Michigan State University

American Educational Research Association
1978 Annual Meeting
Toronto, Ontario, Canada

Printed In USA

ABSTRACT

The purpose of this study was to compare several criterion-referenced reliability coefficients to the Kuder-Richardson estimates and to each other. The Kuder-Richardson formulas 20 and 21, the Livingston, the Subkoviak and two Huynh coefficients (K, $\hat{k}$) were computed for a random sample of 33 criterion-referenced tests. The Subkoviak coefficient yielded the highest mean value; Huynh's Kappa yielded the lowest. The Huynh K and $\hat{k}$ coefficients were highly positively correlated with the Kuder-Richardson 20 and 21 coefficients, and with each other; the Livingston and the Subkoviak indexes were highly correlated with each other. A two-factor principle components solution suggested that only the Subkoviak coefficient measured a test characteristic that differed from the classical (KR) internal-consistency coefficients.

INTRODUCTION

The reliability of a criterion-referenced test can be estimated by several different methods, many derived from differing theoretical frameworks and assumptions. Popham and Husek (1969) and Hambleton and Novick (1973) pointed out that classical estimates of test reliability may be inadequate for tests designed for criterion-referenced interpretations or mastery decisions. Livingston (1972) proposed a criterion-referenced reliability coefficient derived from a redefinition of classical measurement theory in terms of observed-score deviations from the criterion score. The Livingston coefficient subsequently drew criticism from some researchers (Harris, 1971; Hambleton and Novick, 1973). Other researchers (Hambleton and Novick, 1973; Swaminathan, Hambleton, and Algina, 1974) have proposed two-administration consistency-of-mastery-decision indexes as appropriate reliability coefficients for criterion-referenced tests. Still others have presented single-administration indexes of consistency-of-decision reliability (Marshall and Haertel, 1975; Subkoviak, 1977; Huynh, 1977).

Single Administration Reliability Coefficients

The Kuder-Richardson Formula 20 reliability coefficient (Kuder and Richardson, 1937) is given by:

$$r_{xx_{20}} = \frac{K}{K-1}\left(1 - \frac{\Sigma pq}{S_x^2}\right) \qquad (1)$$

where:

K = number of test items
pq = item variance
S_x^2 = test variance

The Kuder-Richardson Formula 21 reliability coefficient (Kuder and Richardson, 1937) is given by:

$$r_{xx_{21}} = \frac{K}{K-1}\left(1 - \frac{\bar{X}(K-\bar{X})}{K S_x^2}\right) \qquad (2)$$

where:

$\bar{X}$ = test mean

Livingston (1972) derived a criterion-referenced reliability coefficient as a correction of classical reliability:

$$r_{cc} = \frac{r_{xx} S_x^2 + (\bar{X} - C)^2}{S_x^2 + (\bar{X} - C)^2} \qquad (3)$$

where:

r_{xx} = classical test reliability

C = criterion score

It should be noted that if $\bar{X} = C$, r_{cc} will, of course, be equal to r_{xx}.

The Subkoviak group coefficient of agreement (Subkoviak, 1977) is the mean of the individual coefficients of agreement (p^i):

$$P_c = \frac{\Sigma\, p_c^i}{N} \qquad (4)$$

where:

p_c^i = the coefficient of agreement[1] for person i.

Huynh (1977) provides two consistency of classification indexes:

---

[1] $\hat{p}^i = r_{xx_{21}}\left(\frac{Xi}{K}\right) + (1 - r_{xx_{21}})\left(\frac{\bar{X}}{K}\right)$

$$K = (p_{11} - p_1^2) / (p_1 - p_1^2) \qquad (5)$$

where[2]:

$$p_{11} = \Sigma f(X, Y)$$

$$p_1 = \Sigma f(X)$$

and, when K > 10:

$$\hat{k} = (p_{oo} - p_o^2) / (p_o - p_o^2)$$

where[3]:

$$p_{oo} = \Sigma f(X, Y)$$

$$p_o = \Sigma f(X)$$

OBJECTIVES OF THIS STUDY

This study proposed to compare empirically the six single-administration reliability coefficients for their usefulness in criterion-referenced testing.

1. The Kuder-Richardson 20 coefficient
2. The Kuder-Richardson 21 coefficient
3. The Livingston (r_{cc}) coefficient
4. The Subkoviak (P_c) Group Index of Consistency
5. The Huynh Index of Consistency (K)
6. The Huynh Index of Consistency ($\hat{k}$)

---

[2]Huynh presents formulas for the estimation of p_{11} and p_1 based on a beta-binominal model.

[3]p_{oo} and p_o are evaluated by reference to univariate and bivariate normal tables, after Gupta (1963).

Data Source

The data for this study were 33 achievement examinations; these examinations represent a random sample of objective format (three-to-five option multiple-choice) criterion-referenced (mastery) examinations from undergraduate teacher education classes, medical school classes and state-wide assessment tests.

The number of examination items ranged from 5 to 143, with a mean of 38.9 items. The number of subjects taking these tests ranged from 5 to 1110 with a mean of 209.9. These examinations had standard deviations from 0.98 to 12.19. Average item difficulty was .26 (proportion incorrect) with a range of .12 to .47 while average item discrimination (D) was .24 with a range of .14 to .47.

Methods

Each of the six reliability coefficients was computed for each examination. The Kuder-Richardson 20 reliability coefficient (and other standard item analysis data) were computed for these exams by standard scoring routines. All other reliability coefficients were computed by a special computer program written for this study.[4] A correlational analysis was then carried out on these data; additionally, a factor analysis was performed to determine the uniqueness of coefficient contribution to the total variance.

Results

Table 1 shows that the Huynh K has the lowest absolute numerical mean value, while Subkoviak's P_c has the highest numerical mean value. Table 2 presents the inter-correlation of these reliability coefficients.

---

[4]The authors wish to acknowledge Mary Yuen for programming assistance.

Table 1

Reliability Coefficients for 33 Tests

Coefficient	Mean	S.D.	Range
KR20	.544	.226	.195 to .923
KR21	.350	.359	-.358 to .870
Kappa (K)	.241	.257	-.177 to .680
Kappa Estimate ($\hat{k}$)	.322	.205	.063 to .644
r_{cc}	.605	.275	-.064 to .900
P_c	.802	.108	.580 to .975

Table 2

Correlation Matrix of Coefficients

	KR20	KR21	K	$\hat{k}$	r_{cc}	P_c
KR20	1.0					
KR21	.949	1.0				
K	.961	.978	1.0			
$\hat{k}$	.974	.996	.993	1.0		
r_{cc}	.693	.644	.590	.562	1.0	
P_c	.393	.298	.226	.147*	.833	1.0

*NS; all others, $p \leq .10$

A principle components factor solution and an orthogonally rotated factor solution are presented in Tables 3 and 4. Factor 1 (unrotated) accounts for 73.4 percent of the total variance while the second factor accounts for 25.4 percent of the variance.

Table 3

Principle Factor Solution

n = 33 Exams

	Factor 1	Factor 2
KR20	.9728	-.1075
KR21	.9761	-.2026
K	.9471	-.3120
$\hat{k}$	.9782	-.2084
r_{cc}	.7162	.6794
P_c	.3466	.9090

Table 4

Varimax-Rotated Factor Solution

n = 33 Exams

	Factor 1	Factor 2
KR20	.9442	.2575
KR21	.9823	.1702
K	.9955	.0578
$\hat{k}$	.9864	.1656
r_{cc}	.4165	.8950
P_c	-.0116	.9727

The principle components analysis suggests that the Huynh coefficients (K and $\hat{k}$) and the Livingston Coefficient (r_{cc}) share a great deal in common with the classical Kuder-Richardson internal consistency reliability estimates, while the Subkoviak coefficient appears to be indexing a test quality that differs from the Livingston/Huynh/Kuder-Richardson formulas.

Discussion

Since all coefficients except the Subkoviak P_c load on the (unrotated) internal consistency factor in this study, it appears that for criterion-referenced tests like those in this sample, it would make little difference whether one uses a classical reliability estimate, the Huynh indexes, or the Livingston coefficient to assess overall examination quality. All of these coefficients appear to be indexing very similar test qualities and one, therefore, has a basis for arguing that the classical coefficients are as appropriate for criterion-referenced tests as the Huynh or Livingston criterion-referenced indexes.

The Livingston Coefficient, r_{cc}, derived directly from classical test theory, loads on both factors and loads highest in the Varimax solution on the Subkoviak factor. This suggests that the Livingston coefficient may be intermediate to the classical and The criterion-referenced coefficients. This result also suggests that the Livingston r_{cc} may be more useful for criterion-referenced reliability than its critics have allowed.

The Subkoviak coefficient does seem to be indexing a test attribute different from the other coefficients considered. Therefore, it may be necessary and desirable to compute both an internal consistency reliability estimate (or K) and the Subkoviak P_c or the Livingston r_{cc} for criterion-referenced tests. Yet, these results do support the usefulness of the

familiar Kuder-Richardson formula 21 coefficient with criterion-referenced examinations. This finding may encourage the criterion-referenced examination user who does not have access to sophisticated computer facilities.

Further Research

The generalizability of the findings in this study is possibly limited by the heterogeneity of the tests in this sample with respect to test length. Further research is indicated, with larger samples of tests that are more homogeneous with respect to test length. Additionally, it is important to attempt to replicate these results for various homogeneous ranges of item difficulties and discriminations.

References

Gupta, S.S. "Probability integrals of multivariate normal and multivariate T." Annals of Mathematical Statistics. 34:792-828, 1963.

Hambleton, R.K., and Novick, M.R. "Toward an integration of theory and methods for criterion-referenced tests." Journal of Educational Measurement. 10:159-170, 1973.

Harris, C.W. "An interpretation of Livingston's reliability coefficient for criterion-referenced tests." Journal of Educational Measurement. 9:27-29, 1972.

Huynh, H. "On consistency of decisions in criterion-referenced testing." Journal of Educational Measurement. 13:253-264, 1977.

Kuder, G.F. and Richardson, M.W. "The theory of the estimation of test reliability." Psychometrika. 2:151-160, 1937.

Livingston, S.A. "Criterion-referenced applications of classical test theory." Journal of Educational Measurement. 9:13-21, 1972.

Marshall, J.L., and Haertel, E.H. "A single-administration reliability index for criterion-referenced tests: The mean split-half coefficient of agreement." AERA Paper: Washington, D.C., 1975.

Popham, W.J. and Husek, T.R. "Implications of criterion-referenced measurement." Journal of Educational Measurement. 6:1-9, 1969.

Subkoviak, M.J. "Estimating reliability from a single-administration of a mastery test." Journal of Educational Measurement. 13:265-276, 1977.

Swaminathan, H., Hambleton, R.K., and Algina, J.J. "Reliability of criterion-referenced tests: A decision-theoretic formulation." Journal of Educational Measurement. 11:263-267, 1974.